補足：統計学の補足・復習

講師：遠山 祐太

最終更新：2024-11-16

はじめに

補足資料の概要

1. 条件付き期待値の復習
2. 推論１：点推定
3. 推論２：仮説検定

条件付き期待値の復習

条件付き期待値

- 計量経済学では、２つの変数 X と Y の関係に関心があることがしばしば。
- 条件付き期待値はその関係を特徴づける方法の一つ。

多変数の確率分布

- X と Y を確率変数とする。
  - X を説明変数、 Y をアウトカム変数としよう。

- 確率分布（probability distribution）
  - 離散版： $x \in \{x_1, \cdots, x_L\}$ および $y \in \{y_1, \cdots, y_L\}$ について、

$$P(y, x) = Prob(X = x, Y = y)$$

  - 連続版（密度関数）： $x \in \mathbb{R}$ および $y \in \mathbb{R}$ について、 $f(y, x)$

条件付き確率分布

- 離散確率変数の確率質量関数（probability mass function）：$P(x) = \sum_{i=1}^{N} P(y_i, x)$ として、

$$P(y|x) = \frac{P(y,x)}{P(x)}$$

- 連続確率変数の密度関数（probability density function）：$f(y|x) = \frac{f(y,x)}{f(x)}$

条件付き期待値

- 離散版：

$$E[Y|X] = \sum_{l=1}^{L} y_l P(Y = y_l|X)$$

- 連続版：

$$E[Y|X] = \int_{-\infty}^{\infty} y f(y|X) dy$$

条件付き期待値の性質

- 命題 CE1：関数 $c(\cdot)$ について、

$$E[c(X)|X] = c(X).$$

- 命題 CE2（線型性）：関数 $a(\cdot)$ および $b(\cdot)$ について、

$$E[a(X)Y + b(X)|X] = a(X)E[Y|X] + b(X).$$

- 命題 CE3： X と Y が独立ならば $E[Y|X] = E[Y]$.

CE3の証明（離散版）

$$
\begin{aligned}
E[Y|X] &= \sum_{l=1}^{L} y_l P(Y = y_l|X) \\
&= \sum_{l=1}^{L} y_l \frac{P(Y = y_l, X)}{P(X)} = \sum_{l=1}^{L} y_l \frac{P(Y = y_l) \times P(X)}{P(X)} = E[Y].
\end{aligned}
$$

ここで、 X と Y の独立性により $P(Y = y, X = x) = P(X = x)P(Y = y)$ を用いた。

くり返し期待値の法則

- 命題 CE4：（**くり返し期待値の法則 ; law of iterated expectation**）

$$E[Y] = E[E[Y|X]]$$

- 条件付き期待値 $E[Y|X]$ の期待値は、条件のない期待値 $E[Y]$ になる。

証明（離散版）

$$
\begin{aligned}
E[Y] &= \sum_{l=1}^{L} y_l P(y_l) \\
&= \sum_{l=1}^{L} y_l \left[\sum_{l'=1}^{L} P(y_l, x_{l'}) \right] \\
&= \sum_{l=1}^{L} y_l \left[\sum_{l'=1}^{L} P(y_l|x_{l'}) P(x_{l'}) \right] \\
&= \sum_{l'=1}^{L} P(x_{l'}) \underbrace{\left[\sum_{l=1}^{L} P(y_l|x_{l'}) y_l \right]}_{E[Y|X]} = E[E[Y|X]]
\end{aligned}
$$

- 連続の場合にも同様に証明できるが、積分の交換のために技術的条件が必要（フビニの定理）

繰り返し期待値法則の発展

- 命題 CE4'：

$$E[Y|X] = E\left[E[Y|X,Z]|X\right]$$

  - X で条件づけても、くり返し期待値の法則は適用できる。

- 命題 CE5： $E[Y|X] = E[Y]$ ならば $Cov(X,Y) = 0$ である。
  - $E[Y|X] = E[Y]$ を **平均独立（mean independence）**とよぶ。
  - 逆は成り立たない。

他の便利な性質

- 離散確率変数 X に対する条件付き確率の法則

$$P(Y) = \sum_{l=1}^{L} P(Y|x_l)P(x_l)$$

- 全分散の法則 (law of total variance)

$$Var(Y) = E[V(Y|X)] + V[E(Y|X)]$$

推論その１：推定

推論の概要

- ここまでは処置効果パラメータの識別について見てきた。
- 実際には、人々（データ）のサンプルがあり、それを未知パラメータの推定に使う。
- RCTにおける統計的推論（statistical inference）について説明する。
  - 点推定（point estimation）
  - 仮説検定（hypothesis testing）

用語：推定対象・推定量・推定値

- 推定対象 (estimand): 推定したいと考えている、母集団の特徴（例：平均）
- 推定量 (estimator)：データ（サンプル）が与えられたときに、目標とするパラメタを推定するルール。いわば、データから推定したい対象への関数。（例：標本平均）
- 推定値 (estimate): データを推定量に代入したときに得られる値。（例：データに基づいて計算した標本平均の値）

ATTの推定

- 推定対象（estimand）のATT

$$E[Y_{1i} - Y_{0i}|D_i = 1] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0].$$

- **条件付き標本平均（conditional sample mean）**を条件付き期待値の推定量（estimator）として用いる。

$$\hat{E}[Y_i|D_i = 1] = \frac{1}{N_1}\sum_{i=1}^{N} Y_i \cdot \mathbf{1}\{D_i = 1\} = \frac{\frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \mathbf{1}\{D_i = 1\}}{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{D_i = 1\}}$$

ATTの推定量

- 標本平均の差分がATTの推定量である。

$$\hat{ATT} = \frac{\frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \mathbf{1}\{D_i = 1\}}{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{D_i = 1\}} - \frac{\frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \mathbf{1}\{D_i = 0\}}{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{D_i = 0\}}$$

- 問：この推定量はなぜATTの良い推定量か？

代替案：線形回帰

- 共変量 X_i と一緒に、処置 D_i に Y_i を線形回帰してもよい。

$$Y_i = \beta_0 + \beta_1 D_i + \beta' X_i + \epsilon_i$$

推定量の性質

$\hat{\mu}_N$ を、未知パラメータ μ の推定量としよう。

1. **不偏性（unbiasednes）**：推定量の期待値が真のパラメータと一致する。

$$E[\hat{\mu}_N] = \mu$$

2. **一致性（consistency）**：推定量が真のパラメータに確率収束（converge in probability）する。

$$\forall \varepsilon > 0, \quad \lim_{N \to \infty} \ Prob(|\hat{\mu}_N - \mu| < \varepsilon) = 1$$

- 直感：サンプルサイズが大きくなればなるほど、推定量と真のパラメータは確率 1 で近づく。
- 注意：数列の収束（convergence in sequence）とは少し違う。

既出の推定量を再考

- **大数の法則（law of large numbers）**：標本平均は母平均に確率収束する。

$$\frac{1}{N}\sum_{i=1}^{N} X_i \xrightarrow{p} E[X]$$

- 連続写像定理（continuous mapping theorem）を用いると、上の推定量にも適用できる。

$$\frac{\frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \mathbf{1}\{D_i = 1\}}{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{D_i = 1\}} \xrightarrow{p} \frac{E[Y_i D_i]}{E[D_i]} = E[Y_i | D_i = 1]$$

- 練習問題：上の等式部分を示せ。（ヒント：law of iterated expectationを用いる）

推論その２：仮説検定

仮説検定

- 検定（testing）：母集団のパラメータに関する仮説の真偽を標本を使って判断する
- 例１：母集団の平均年齢は45歳か？
- 例２：母集団で男女の試験の点数は異なるか？
- 課題：標本統計量（sample statistics）はランダムなので、
  - ただのランダムな現象なのか
  - 母集団における真の効果（違い）なのか

  を区別しなければいけない。

母平均の例

1. 標本平均 $\bar{Y}$ を計算する。
2. **帰無仮説（null hypothesis）**と**対立仮説（alternative hypothesis）**を決める。ある μ について
   - 帰無仮説： $H_0 : E[Y] = \mu$
   - 対立仮説： $H_1 : E[Y] \neq \mu$
3. 帰無仮説 H_0 が正しいと仮定すると、標本平均 $\bar{Y}$ は μ に近い値になるはずである。
4. もし標本平均 $\bar{Y}$ が μ から「あまりにかけ離れて」いる場合、**帰無仮説 H_0 は棄却（reject）される。**（背理法）

- 問い：「あまりにかけ離れている」ことをどう決定するか？

予備知識：標準誤差

- $V(\bar{Y})$ を標本平均の（母）分散としよう。
- もし Y_i が**独立同一分布（independent and identically distributed; i.i.d.）**ならば

$$V(\bar{Y}) = \frac{1}{N^2} \sum_{i=1}^{N} V(Y_i) = \frac{V(Y)}{N}.$$

- **標準誤差（standard error; SE）**：標本平均の標準偏差

$$SE(\bar{Y}) = \sqrt{V(Y)/N}$$

標準誤差の推定値

- 母分散 $V(Y)$ は未知パラメタであるため、それをサンプルから計算可能な不偏分散 $S(Y)$ で置き換えた、標準誤差の**推定値（estimate）**を用いる。

$$\hat{SE}(\bar{Y}) = \sqrt{\hat{V}(Y)/N},$$

ただし $\hat{V}(Y) = S(Y) = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$.

t統計量

- 帰無仮説 $H_0 : E[Y] = \mu$ を考えよう。
- **t統計量（t-statistics）**を次のように定義する。

$$t(\mu) = \frac{\bar{Y} - \mu}{\hat{SE}(\bar{Y})}$$

- 帰無仮説が正しいならば、 $t(\mu)$ はとある分布に従う。
- もし $t(\mu)$ の実現値がその分布に従いそうになければ、帰無仮説を棄却する。
- 問：どんな分布だろうか？

中心極限定理（central limit theorem; CLT）

- 平均 μ 、分散 σ^2 の確率変数 Y からi.i.dな標本 $Y_1, \ldots, Y_N$ を取り出す。このとき、次の Z は正規分布に分布収束（converge in distribution）する。

$$Z = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{Y_i - \mu}{\sigma} \xrightarrow{d} N(0,1)$$

- この定理より

$$t(\mu) = \frac{\bar{Y} - \mu}{\hat{SE}(\bar{Y})} = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i - \mu}{\sqrt{\hat{V}(Y)/N}} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{Y_i - \mu}{\sqrt{\hat{V}(Y)}} \overset{approx}{\sim} N(0,1)$$

Rによる中心極限定理のシミュレーション

- 母数（確率） $p = 0.4$ のベルヌーイ分布（*Bernoulli distribution*）に従う確率変数 Y_i を考える。
- ここで、 $E[Y] = 0.4$ かつ $V[Y] = 0.4 \times (1 - 0.4)$.
- Z を次のように定義する。

$$Z = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{Y_i - E(Y)}{\sqrt{V(Y)}}$$

- N が大きくなればなるほど、 Z の分布が正規分布に近づくことを確かめよう。

関数定義

- この関数では、ベルヌーイ分布から`samplesize`の観測を取り出して各標本について Z を計算する手続きを`Nreps`回繰り返す。

```
f_simu_CLT <- function(Nreps, samplesize, distp ){

  output = numeric(Nreps)
  for (i in 1:Nreps ){

    test <- rbinom(n = samplesize, size = 1, prob = distp)

    EY <- distp
    VY <- (1 - distp)*distp

    output[i] <- ( mean(test) - EY ) / sqrt( VY / samplesize )
  }

  return(output)

}
```

```r
# シード値を設定
set.seed(12345)

# シミュレート
Nreps = 500
result_CLT1 <- f_simu_CLT(Nreps, samplesize = 10 , distp = 0.4 )
result_CLT2 <- f_simu_CLT(Nreps, samplesize = 1000, distp = 0.4 )

# 比較のため正規分布からの観測取り出し
result_stdnorm = rnorm(Nreps)

# データフレーム作成
result_CLT_data <- data.frame(  Ybar_standardized_10 = result_CLT1,
                              Ybar_standardized_1000 = result_CLT2,
                              StandardNormal = result_stdnorm )
```

- 分布を図示しよう。

```
# tidyverse読み込み
library("tidyverse")

# meltを使ってresult_dataのフォーマットを変更
data_for_plot <- tidyr::pivot_longer(data = result_CLT_data, cols = everything())

# ggplot2で図示
fig <-
  ggplot(data = data_for_plot) +
  xlab("Sample mean") +
  geom_density(aes(x = value, colour = name ),  ) +
  geom_vline(xintercept=0 ,colour="black")
```

```
plot(fig)
```

- N が大きくなるにつれ、正規分布に近づく。

中心極限定理に基づく仮説検定

- 標準正規分布は $\mu = 0$ かつ $\sigma = 1$
- この分布では、絶対値が2より大きい値をとるのはたった約5%にすぎない！
- もし $t(\mu)$ の絶対値が2より大きければ、5%の水準で帰無仮説が正しいとはいえないと判断する。
- 標本平均が「有意に」0とは異なるともいう。

2グループ間における標本平均の差の検定

- 処置効果が0かそうでないかを検定したい。
- 帰無仮説は

$$H_0 : E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = 0.$$

- この場合のt検定量は

$$t = \frac{\bar{Y}_1 - \bar{Y}_0}{\hat{SE}(\bar{Y}_1 - \bar{Y}_0)}.$$

- ここで、 $\bar{Y}_d$ はグループ d の条件付き標本平均である。

- 標準誤差は

$$SE(\bar{Y}_1 - \bar{Y}_0) = \sqrt{\frac{V^1(Y)}{N_1} + \frac{V^0(Y)}{N_0}}$$

ただし $V^d(Y)$ はグループ d の観測の母分散である。